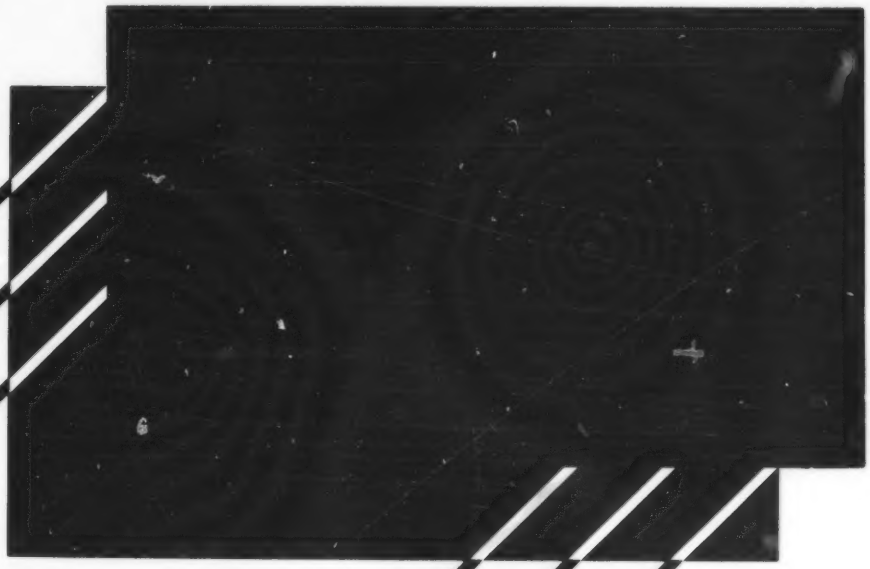
Gouvernement du Canada
Ministère des Communications

Government of Canada
Department of Communications

Le Centre canadien de recherche sur l'informatisation du travail
Canadian Workplace Automation Research Centre



Canada

MACHINE TRANSLATABILITY OF AERONAUTICAL INFORMATION PUBLICATION COPSYS MANUALS CENSUS AND INTERCENSAL STUDIES

by

John Lehrberger

January 1990

Cat. no. Co 28-1/55-1990 E

ISBN 0-662-/7915-3

92-02867/

TABLE OF CONTENTS

		Page
1.	INTRODUCTION	1
1.1	Purpose of this study	1
1.2	Text domains	1
1.3	Methodology	1
2.	AERONAUTICAL INFORMATION PUBLICATION	6
2.1	The corpus	6
2.2	Size of vocabulary	7
2.3	Sematic range	12
2.4	Syntactic properties	13
2.5	Conclusion	16
3.	CORPORATE OPERATING SYSTEMS (COPSYS)	17
3.1	The corpus	17
3.2	Size of vocabulary	17
3.3	Sematic range	21
3.4	Syntactic properties	22
3.5	Conclusion	27
4.	CENSUS AND INTERCENSAL STUDIES	28
4.1	The corpus	28
4.2	Size of vocabulary	29
4.3	Semantic range	33
4.4	Syntactic properties	34
4.5	Conclusion	37

1. INTRODUCTION

1.1 Purpose of this study

The purpose of this study is to examine a variety of documents currently translated by departments or agencies of the Canadian government in order to determine their suitability for machine translation. Sample texts provided by the departments or agencies concerned are analyzed in terms of their lexical, grammatical and semantic properties. The overall linguistic complexity of a sample from each text domain serves as a basis for rating that domain as a possible candidate for machine translation.

1.2 Text domains

Sample texts have been furnished by Transport Canada, Canada Post and the Department of Defense. The particular text domains are as follows:

Transport Canada: Aeronautical Information Publication (A.I.P.)
Size of sample - 11,155 words

Canada Post: Corporate Operating Systems (COPSYS)
Size of sample - 9,724 words

Statistics Canada: Census and Intercensal Studies (CENSUS)
Size of sample - 12,009 words

1.3 Methodology

Certain aspects of linguistic complexity are effective indicators of the tractability of texts for the purpose of machine translation. These include size of vocabulary, semantic range and syntactic complexity.

(i) **Size of vocabulary.** The term *vocabulary*, as used here, corresponds closely to *dictionary* or *lexicon* in an MT system. Since most MT systems perform some morphological analysis, the dictionary usually lists only base forms of words (for certain classes of words) without all the inflectional variants. For example, the verb *fly* may be listed, but not *flies, flew, flying, flown*, or the noun *airport*, but not *airports*. However, the software used here for vocabulary study (FATRAS - Full Text Retrieval System) provides a list of *forms* including all the inflectional variants. Abbreviations and alphabetical designators are also treated as individual forms, but numbers are not. All numbers having the same number of digits are taken as instances of the same form. Thus all one digit numbers are represented by 0 in FATRAS' vocabulary list, all two digit numbers by 00, all three digit numbers by 000, etc. This reflects the fact that the presence of a great many numerical expressions in a text does not in itself add significantly to the syntactic or semantic complexity of the text.

A base form together with its inflectional variants will be referred to as a *lexical lemma* (or simply *lemma*), as will the set of all numbers having n digits, where n is a positive integer. In addition to verbs and nouns, as illustrated above, adjectives together with their comparative and superlative forms will be taken as lemmas (eg, *great, greater, greatest*). A vocabulary is then understood as a set of lemmas, in the sense just defined.

In order to estimate the size of vocabulary required for all texts in a given domain, a representative corpus is chosen, its vocabulary is determined with the help of FATRAS, and a projection is made to the larger set of texts - those in existence as well as those that may appear in the domain in the near future. The basis for this projection is the *lexical growth rate* observed in the corpus; ie, the rate at which new lemmas are introduced as more and more text is taken into account (a new lemma is "introduced" if any member of that lemma appears in the added text). The procedure is to divide the corpus into a number of roughly equal segments and to determine at the end of each segment the number of lemmas introduced and the number of words (ie, word tokens) of text up to that point. The lexical growth rate for a given segment is the **ratio** of the number of lemmas introduced in that segment to the number of words of text in the segment. The manner in which this ratio changes over successive segments gives an indication of the vocabulary size for the text domain. The

growth rate is expected to approach zero if enough text is taken into account, the amount depending on the domain.

If the first n segments of a corpus introduce all the lemmas needed for the whole domain, the growth rate is zero for the remaining segments. (Of course, this is not likely to happen; normally the corpus vocabulary is a proper subset of the domain vocabulary.) But the converse does not hold: if the growth rate falls to zero after n segments, it is not necessarily the case that all the lemmas needed for the domain have been introduced up to that point. There may be some less frequently used words that have not yet shown up or, more importantly, there may be a large number of words that occur frequently in certain subdomains, but rarely, if at all, elsewhere. A "representative" corpus would tap all these subdomains in just the right proportions; in practice that ideal is seldom attained. Also, new material is constantly being published, possibly introducing new vocabulary items into the domain. This is especially true of rapidly developing technical domains.

On moving from one subdomain to another, from one section of a text to another in the corpus, there may be a sharp increase in the value of the lexical growth rate obtained just before that point. This was the case in the COMPUT domain[1] where the value of the lexical growth rate increased dramatically in moving from a segment of the corpus representing one type of computer system manual to a segment from another type. Since the lexical growth rate does not, in general, approach the zero value in a smooth manner, it is necessary to be on the lookout for possible sources of such changes. The result is that although we may be able to predict with confidence that the vocabulary for a certain domain contains in excess of k lemmas, we may not be able to specify the amount in excess with any degree of precision.

One further note. In those cases where there are sharp increases in lexical growth rate in some segments of the corpus, it should not be assumed that the increases correspond to changes in subdomains. There are other sources such as lists of place names, abbreviations, and parts lists. Such increases in vocabulary may not add significantly to the overall linguistic

---

[1] See "The Machine Translatability of Computer System Manuals", J. Lehrberger, CWARC, 1988.

complexity, and this must also be taken into account in assessing the effect of vocabulary size on the tractability of the text for the purpose of MT.

(ii) **Semantic range.** The success of MT still depends heavily on its application to texts with restricted subject matter. Restricting the semantic range reduces the degree of polysemy in the vocabulary and makes it easier to deal with the homography problem, a very serious impediment to machine translation. Nevertheless, even a restricted subject matter may include many subtopics; the number and variety of these must be taken into account, and the extent to which they overlap or are otherwise related.

(iii) **Syntactic properties.** Rather than trying to apply some general definition of syntactic complexity to the sample texts in order to determine their tractability, the emphasis here is on the identification of particular syntactic properties that are known to make a text difficult to parse. The list of structures given below is used for that purpose. (Sentence length is included in the list since the general tendency is that longer sentences are more difficult to parse.)

(A) SENTENCE TYPES

declarative, interrogative, imperative

(B) TOPICALIZATION

passive, extraposition, cleft and pseudo-cleft, inversion

(C) CLAUSAL STRUCTURE

relative clause (full and reduced); clause introduced by subordinate conjunction; clause as complement of verb, adjective or noun; abbreviated clause; appositive clause; clause with subjunctive verb; comparative clause; *Ving, Ved, to V* adverbial clauses and postnominal modifiers not derived from relative clauses

(D) COORDINATION

conjunction by *and, or*

(E) ELLIPSIS

(F) TELEGRAPHIC SENTENCES

(G) NOUN STACKING
complex nominal with string of 2 or more nouns, possibly interspersed with adjectives and possibly including one or more coordinate conjunctions

(H) PARENTHETICALS
types, distribution, kinds of expressions enclosed

(I) TENSE

(J) MODALS AND SEMI-AUXILIARIES

(K) SENTENCE LENGTH

(L) TEXT STRUCTURE
sentence linking, information structure, text formatting

FINDINGS

2. AERONAUTICAL INFORMATION PUBLICATION (A.I.P. Canada)

This domain will be referred to as A.I.P. The publication consists of pre-flight information for users of Canadian airspace:

> "It provides flight crews with a single source for information concerning rules of the air and procedures for aircraft operation in Canadian airspace" and "includes those Air Regulations and Air Navigation Orders of interest to pilots and flight navigators."

The publication is updated every 56 days with changes affecting 50 to 150 pages. The new material usually constitutes 20% to 40% of the page. Turnaround time for translation is about 2 weeks.

2.1 The corpus

The part of A.I.P. Canada submitted to detailed analysis is a section from Rules of the Air and Air Traffic Services entitled SPECIAL PROCEDURES. It contains 11,155 words of text, as counted by FATRAS, and includes the following topics:

- Air Traffic Flow Management
- Advanced Flow Control Procedures
- Domestic Clearance - North Atlantic
- North Atlantic Oceanic Control Procedures
- North Atlantic Minimum Navigation Performance Specification Air-space
- North American Routes for North Atlantic Traffic
- Required Navigation Performance Capability Airspace
- Canadian Minimum Navigation Performance Specifications Airspace and Certification
- Transition Canadian Domestic - CMNPS Transition Area
- Commonly Used Routes - Winnipeg/Montreal FIRs
- Fuel Conservation High Level Airspace
- Trans-Oceanic Flight - General Aviation Aircraft
- Photographic Survey Flights

2.2 Size of vocabulary

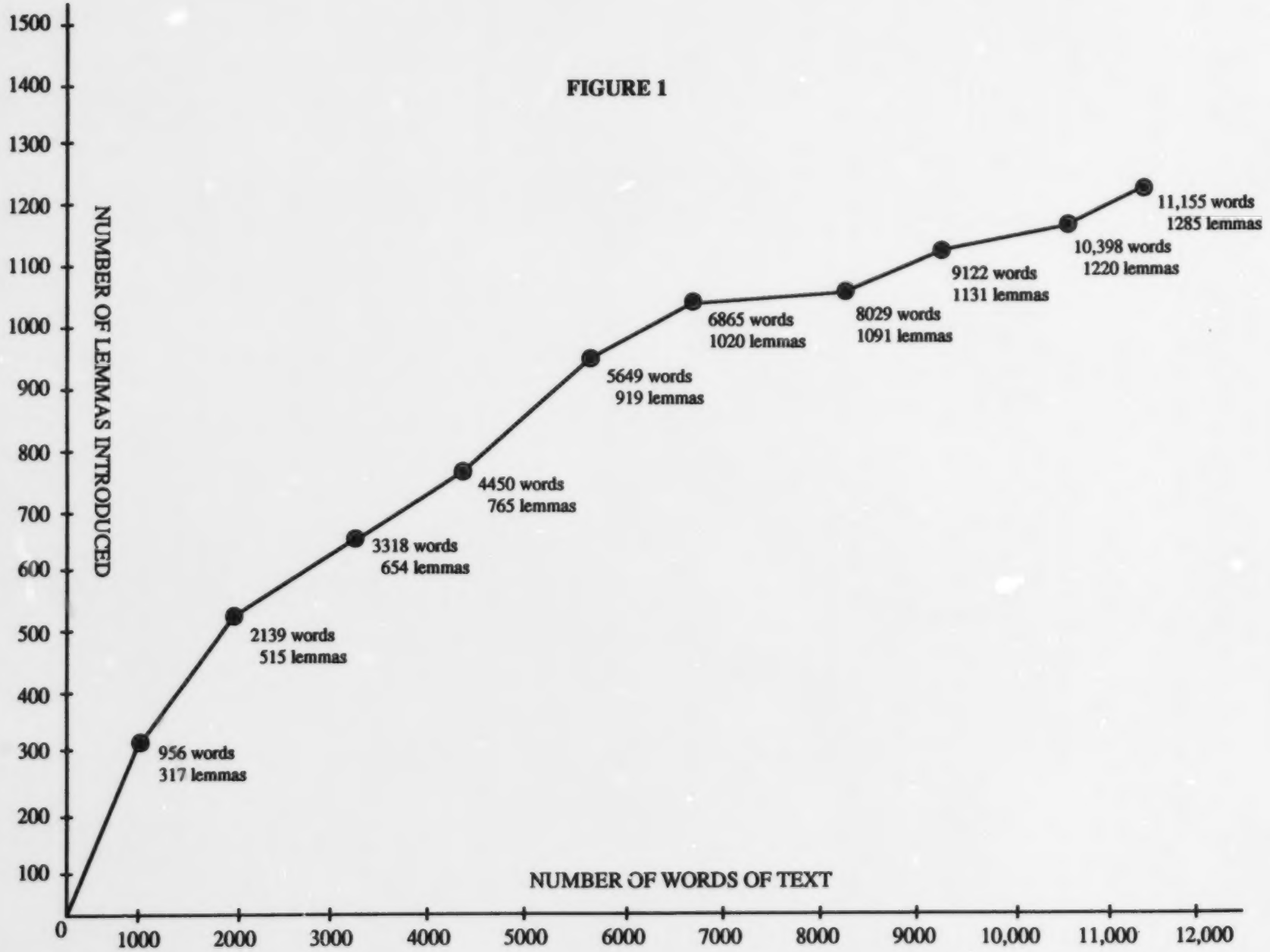
The 11,155 word corpus was divided into 10 segments. Table 1 shows the number of words in each segment, the number of lexical lemmas introduced there and the lexical growth rate for each segment. The lexical growth rate is expressed, for convenience, as the number of lemmas introduced per 1,000 words of text, ie, (lemmas/words) x 1,000. In Figure 1, page 9, the changes in lexical growth rate are shown graphically; this is referred to as the lexical growth rate curve. The units on the horizontal axis represent the number of words of text and those on the vertical axis represent the vocabulary count (number of lemmas introduced). Successive differences in these values yield the values in the second and third columns of Table 1. The slope of the curve for any segment represents the growth rate for that segment.

SEGMENT	WORDS OF TEXT	LEMMAS INTRODUCED	LEXICAL GROWTH RATE
1	956	317	332
2	1,183	198	167
3	1,179	139	118
4	1,132	111	98
5	1,199	154	128
6	1,216	101	83
7	1,164	71	61
8	1,093	40	37
9	1,276	89	70
10	757	65	86

TABLE 1

The growth rate decreases until the fifth segment, where it shows a 31% increase. This increase can be attributed in part to the fact that the fifth segment of the corpus contains 2 lists of air routes, which introduce a large number of place names (Akraberg, Churchill, Machrihanish,etc.). The next 3 segments show a steady decrease to 37 new lemmas per 1,000 words of text. If our corpus had ended at that point (9,122 words) we would see a curve which is beginning to level off and we might expect the growth rate to fall steadily to near zero, assuming that trend to continue. However, in the ninth segment the value almost doubles, followed by a 23% increase in the tenth segment. Place names and designators of communications stations are responsible for much of the increase in segment 9, but this is not the case in segment 10.

FIGURE 1


NUMBER OF LEMMAS INTRODUCED
1500
1400
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
0
956 words
317 lemmas
2139 words
515 lemmas
3318 words
654 lemmas
4450 words
765 lemmas
5649 words
919 lemmas
6865 words
1020 lemmas
8029 words
1091 lemmas
9122 words
1131 lemmas
10,398 words
1220 lemmas
11,155 words
1285 lemmas
NUMBER OF WORDS OF TEXT
1000
2000
3000
4000
5000
6000
7000
8000
9000
10,000
11,000
12,000

With a corpus of more than 11,000 words, the lexical growth rate curve has not levelled off and the increase in segment 10 cannot be attributed to the infusion of large numbers of place names, designators or other items which have little effect on overall linguistic complexity.

In Figure 2, page 11, these results are compared with those from the COMPUT domain, where a corpus of 9,914 words was used. A.I.P. shows a slightly larger vocabulary at all points along the curve. There is no data for COMPUT beyond 9,914 words, but in light of the increased growth rate in A.I.P. in that region, its vocabulary would likely remain larger than that of COMPUT if the latter were extended. This is supported by evidence presented in the discussion of the semantic range of A.I.P. in the next section. The vocabulary of the A.I.P. domain can be expected to contain well in excess of 5,000 lemmas.

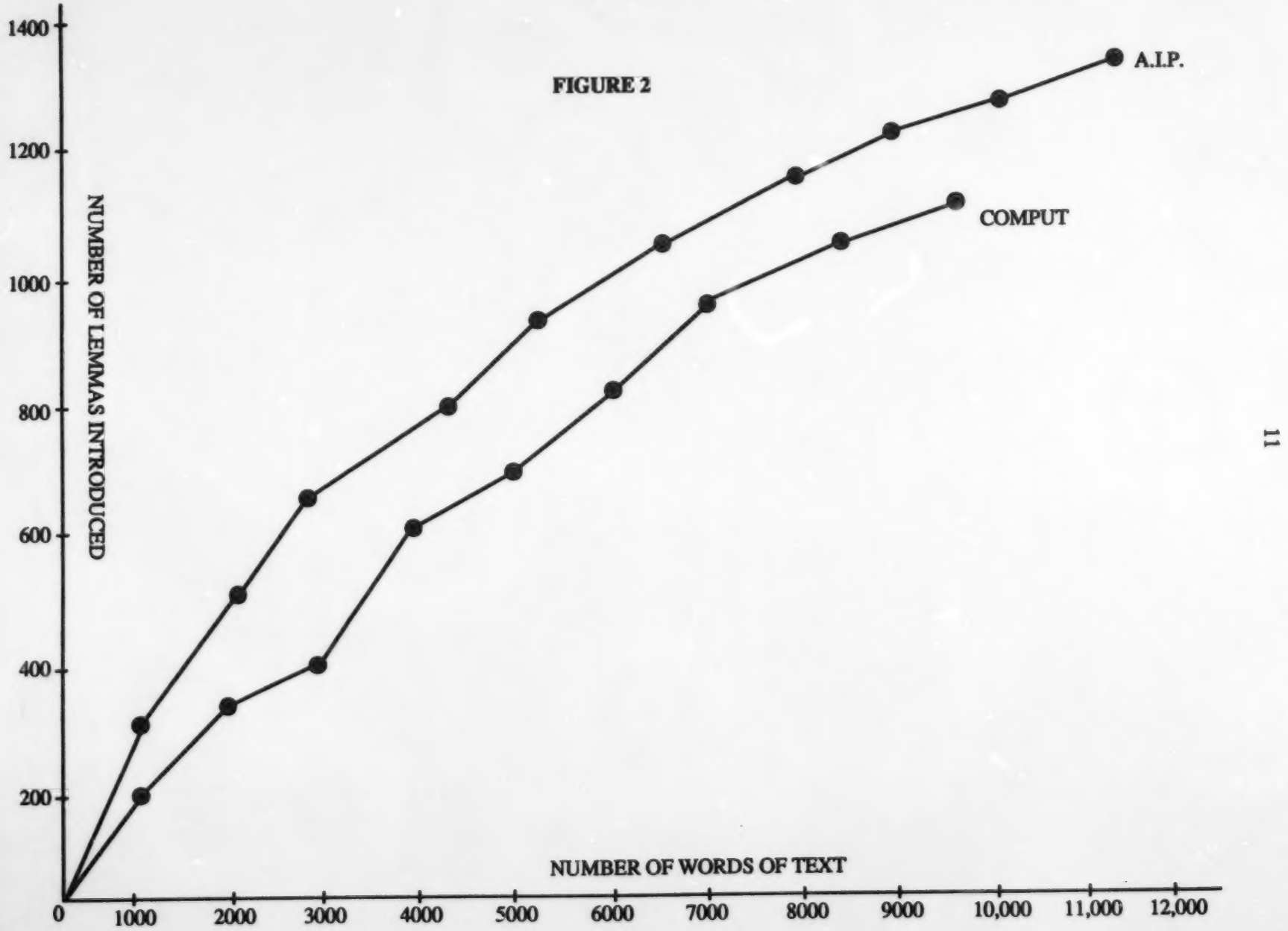


FIGURE 2
A.I.P.
COMPUT
NUMBER OF LEMMAS INTRODUCED
NUMBER OF WORDS OF TEXT
1400
1200
1000
800
600
400
200
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10,000
11,000
12,000

2.3 Semantic range

A.I.P. Canada is divided into 12 sections, as follows:

1. General information (GEN)
2. Aerodromes (AGA0
3. Communications (COM)
4. Meteorology (MET)
5. Rules of the air and air traffic services (RAC)
6. Facilitation (FAL)
7. Search and rescue (SAR)
8. Aeronautical charts and publications (MAP)
9. Licensing, registration and airworthiness (LRA)
10. Airmanship (AIR)
11. Notices to airmen (NOTAM)
12. Aeronautical information circular (AIC)

What these topics have in common is simply the fact that they furnish pre-flight information for flight crews. The semantic diversity of that information is evident from the inclusion of Communications, Meteorology, Licensing and Registration along with Air Traffic Rules and Services.

The largest section, RAC, undergoes the most changes in the 56 day updating. It covers a wide variety of topics in addition to the special procedures discussed in 2.1 above. These topics include: Air Traffic and Advisory Services, Radar Service, Conservation (Fur and Poultry Farms, Migratory Birds, Reindeer, Caribou, Moose, etc.), Flight Planning, Airport Operations, Visual Flight Rules, Instrument Flight rules, etc. Although most of the subdomains show considerable overlap, a few stand apart (eg, conservation). This is also the case in some other sections: within AIR there is a subsection, Medical Facts for Pilots, containing information on fatigue, hypoxia, alcohol, drugs, vertigo, carbon monoxide, vision, middle ear discomfort, etc.; and within LRA there are topics such as Change of Ownership, Importation of Aircraft, Licensing and Registration. It is evident that a sizeable vocabulary is needed to cover all these topics, although some words would be restricted to a single subdomain.

In summary, A.I.P. has a broad semantic range with a fairly large number of subdomains; the unifying factor is not the semantic nature of the information from these subdomains, but the use to which that information is put.

2.4 Syntactic properties

(A) Sentence types

There are no interrogative sentences in the corpus. A few imperatives occur and these are mostly of the type "Call Moncton Centre" or "See GEN 1.0".

(B) Topicalization

The majority of sentences are in the passive. Extraposition is used occasionally. There is one example of inversion ("Should it be determined that flight to the airport of destination is undesirable, the pilot will...") and there are no cleft or pseudo-cleft sentences.

(C) Clausal structure

Sentences tend to be rather complex, with a great number of subordinate clauses throughout the corpus. Relative clauses (especially those with *wh + be* deletion), clauses that occur as complements of verbs, adjectives, etc., and clauses introduced by subordinate conjunctions all occur with very high frequency. Abbreviated clauses are quite common (eg, if required, unless otherwise instructed, as noted, when making position reports, while executing the turnback, when out of range, whether continuing to destination or turning back, whenever possible). There are several occurrences of clauses with subjunctive verb (eg, It is essential that ATS be informed). A variety of other clause types also occur, but those mentioned above are most common.

(D) Coordination

Coordination of noun phrases is very frequent, followed by verb or verb phrase coordination and sentence coordination, in that order. Scope of conjunction is often a problem and occasionally conjuncts are not of the same grammatical category (eg, "information can be obtained by contacting the following:...The shift manager or ATFM unit of the applicable

Canadian ACC for the airport of departure or by contacting the Central Flow Control Facility...").

(E) Ellipsis

Conjunction reduction is about average and there are a few occurrences of gapping(eg, "lattitude shall be expressed in degrees and minutes, longitude in degrees only"). Subjectless sentences are found in list context. *Wh + be* deletion in reduced relative clauses is of very high frequency.

(F) Telegraphic sentences

Unlike the situation in maintenance manuals, instructions in A.I.P. are presented in the form of full sentences, usually employing modal auxiliaries for that purpose. There are a few examples of article deletion suggesting telegraphic style, but this is rare (eg, "When turnback is completed, heading should be adjusted").

(G) Noun stacking

The corpus rates very high in the number of noun phrases that contain strings of 2 or more nouns. Some typical examples are:

(a) arrival delay prediction
(b) high traffic density areas
(c) special aircraft equipment requirements
(d) applicable flight line photographic tracks
(e) Canadian domestic lateral separation minimum
(f) extended range twin-engined aircraft operations
(g) advanced flow control restriction (AFCR) message
(h) Air Traffic Flow Management Programs (ATFM)
(i) Fuel Conservation High Level Airspace
(j) Canadian Minimum Navigation Performance Specifications (CMNPS) Certification
(k) radiotelephony distress or urgency signal

Here we have the usual problems of bracketing the noun strings and deciding which noun the adjective applies to, as well as the interaction with scope of conjunction (k). In some cases capitalization indicates that the noun phrase might be entered in the dictionary as an idiom, especially when the expression is immediately followed by initials in parentheses,

as in (j). But allowing for this possibility, there are still many difficult examples remaining, a large number of which contain more than 2 nouns in sequence.

(H) Parentheticals

The most frequent use of parentheses is to enclose initials, as in (g), (h) and (j) above. Because they are usually placed immediately after the expressions the initials stand for, such parentheticals often occur inside a noun string, as in (g) and (j). But note the exception in (h) where the word *Programs* intervenes.

(I) Tense

Present tense is used almost exclusively, with only 7 occurrences of simple past tense and 17 of present perfect in the corpus.

(J) Modals and semi auxiliaries

The corpus contains the full range of modal auxiliaries and they are used very frequently. The preface to A.I.P. Canada (not included in the corpus) comments on the use of *should* and *shall*: "Throughout the A.I.P. the term "should" implies that the Department of Transport encourages all pilots to conform with the applicable procedure. The term "shall" implies that the applicable procedure is mandatory and supported by regulations or orders." In fact, *should* is also used in the sense of *if*, for example in the sentence quoted above under Topicalization, and the following: "Should the organized track system at the time of the incident extend to the Northern part of the NAT Region, the aircraft concerned may be required to accept a lower than optimum flight level". By using *shall* to indicate that a procedure is mandatory the A.I.P. avoids imperative sentences in many instances. Occurrences of *will* are normally optative rather than indicating future time. There are several occurrences of the semi-auxiliary *be to*.

(K) Sentence length

The average sentence length for the corpus is 23.2 words and 27.4% of the sentences contain 30 or more words. These figures are high.

(L) Text structure

(i) Sentence linking. There are just 4 occurrences of intersentential pronouns in this 11,155 word corpus.

(ii) Information structure. There is no widespread correlation between information and syntactic structure that could be used as a basis for defining information formats over a significant part of the text.

(iii) Text formatting. A noticeable feature of the corpus is the presence of lists of routes and tracks consisting mainly of place names, abbreviations and compass directions. These lists do not appear to present any difficult problems.

2.5 Conclusion

A.I.P. Canada covers a wide range of topics, all of them contributing information of value to flight crews operating in Canadian airspace. Because of the variety of topics, the text domain of A.I.P. consists of a number of different semantic domains from which the relevant information is drawn, rather than a single restricted semantic domain. This results in an extensive vocabulary, with ample opportunity for homography.

Overall, the corpus rates high in syntactic complexity. Sentences tend to be long (average = 23.2 words, compared to 19.5 for the CEIC section of the COMPUT domain) with a large number and variety of subordinate clauses. The passive voice predominates throughout. The rate of occurrence of complex nominals with noun stacking is very high and the noun strings are frequently quite long.

Since A.I.P. has a broad semantic range, an extensive vocabulary, and rates high in syntactic complexity, it is not recommended as a candidate for machine translation.

3. CORPORATE OPERATING SYSTEMS (COPSYS) - CANADA POST

This domain will be referred to as COPSYS. The COPSYS manual contains the approved methods and procedures used at Canada Post to process mail. In fact, there are about 450 COPSYS manuals at Canada Post Work Centres around the country; they have the same general form and content, but may differ from one another in some details, reflecting differences between individual Work Centres. Whenever a change in mail processing takes place at any Work Center, the manual there is revised accordingly, which involves a great many contributors to manuals in different localities. There is, however, a system guide, or "generic" manual, that provides a degree of uniformity in the system.

3.1 The corpus

The corpus consists of 9,661 words of text, as counted by FATRAS, drawn from the COPSYS manual. The contents of the manual are as follows:

- INTRODUCTION
- STRUCTURE AND CONTENT
- OPERATING PROCEDURES
- LAYOUTS
- SORTATION SCHEMATICS OVERVIEW
- MAIL FLOW CHARTS OVERVIEW
- WORD PROCESSING SPECIFICATIONS
- WORDSTAR FILES AND DISKETTE

Not included in the corpus are tables with only numerical entries and long lists of place names. These do not add significantly to the linguistic complexity of the text.

3.2 Size of vocabulary

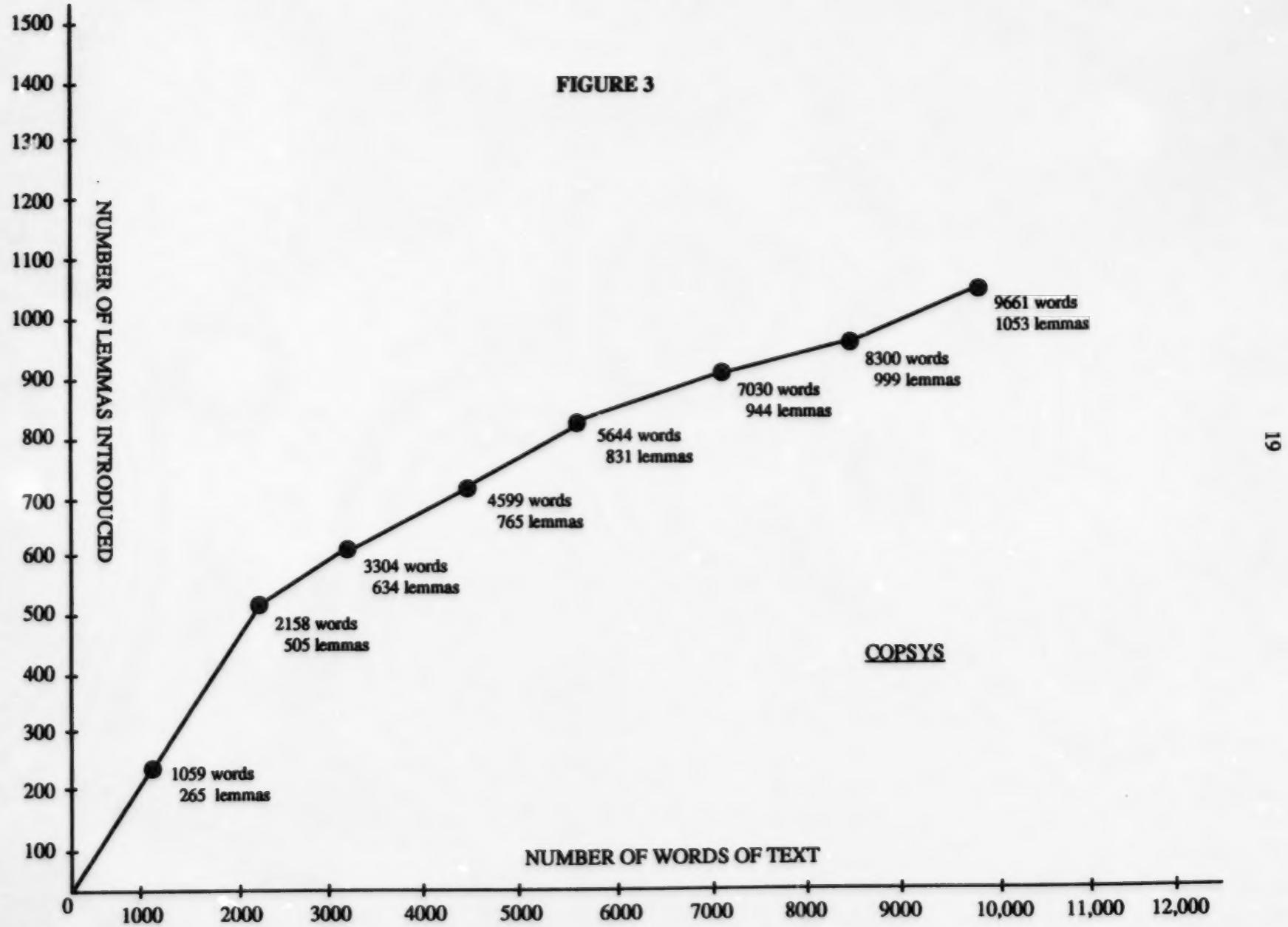
The corpus was divided into eight segments for the vocabulary study. Table 2 shows the number of words in each segment, the number of lexical lemmas introduced there and the

lexical growth rate for each segment (ie, the number of lemmas introduced per thousand words of text). Figure 3, page 19, is the lexical growth rate curve obtained from this data. In Figure 4, page 20, the curve for COPSYS is compared with that for A.I.P.

SEGMENT	WORDS OF TEXT	LEMMAS INTRODUCED	LEXICAL GROWTH RATE
1	1,059	265	250
2	1,099	240	218
3	1,146	129	113
4	1,295	131	101
5	1,045	66	63
6	1,386	113	82
7	1,270	55	43
8	1,361	54	40

TABLE 2

FIGURE 3


NUMBER OF LEMMAS INTRODUCED
1500
1400
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10,000
11,000
12,000
NUMBER OF WORDS OF TEXT
1059 words
265 lemmas
2158 words
505 lemmas
3304 words
634 lemmas
4599 words
765 lemmas
5644 words
831 lemmas
7030 words
944 lemmas
8300 words
999 lemmas
9661 words
1053 lemmas
COPSYS

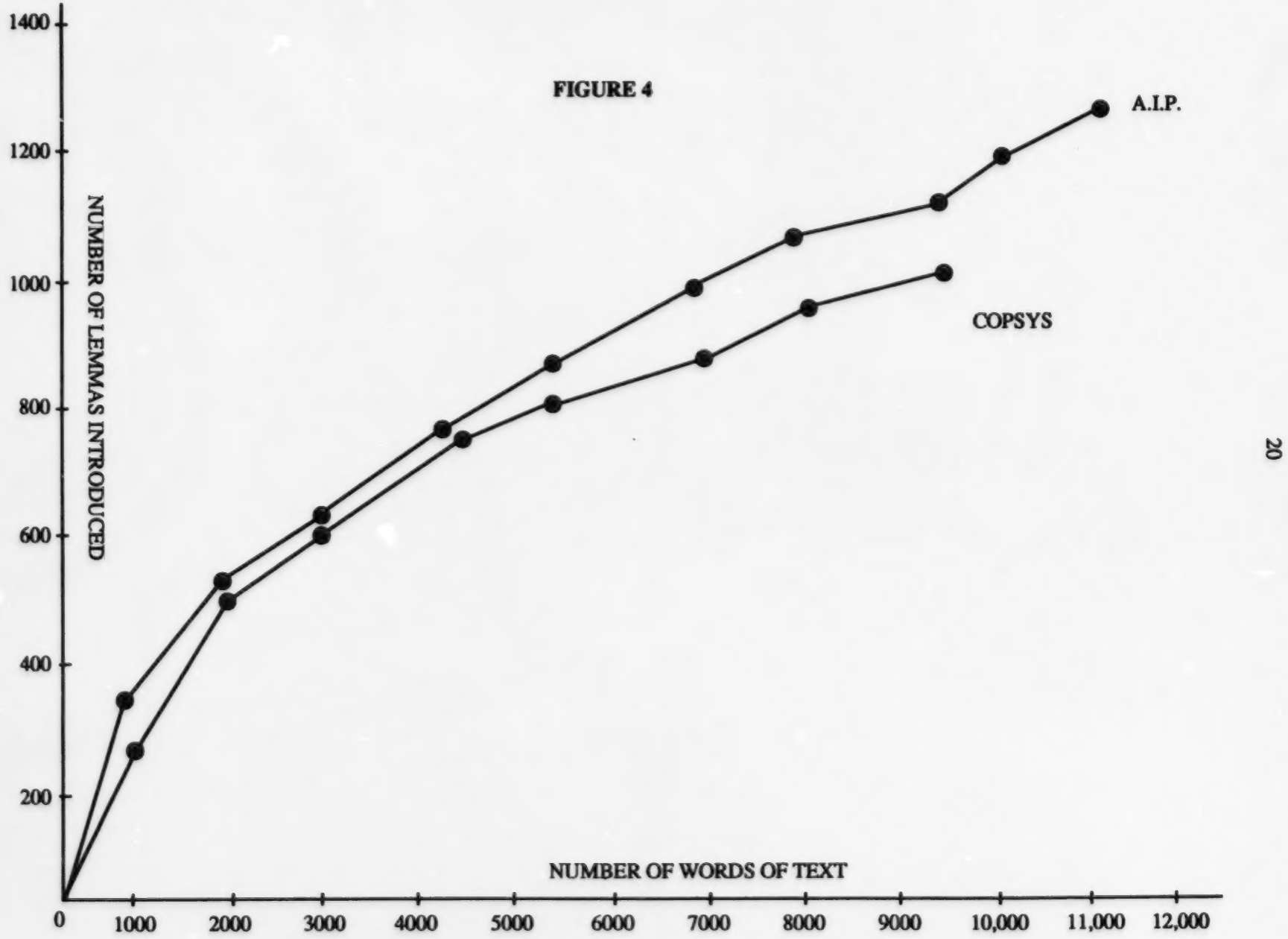


FIGURE 4
A.I.P.
COPSYS
NUMBER OF LEMMAS INTRODUCED
NUMBER OF WORDS OF TEXT
1400
1200
1000
800
600
400
200
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10,000
11,000
12,000

There is a general decrease in the lexical growth rate of COPSYS, except for the sixth segment. This general trend should continue beyond the corpus, since the main topics in COPSYS have already been covered in the sample, and in light of the restricted semantic range (section 3.3), large infusions of new vocabulary would seem unlikely - with the possible exception of place names. It appears then that the vocabulary for the domain can be expected to contain 2,000 - 4,000 lexical lemmas, exclusive of place names.

There is a General Glossary prepared by Linguistic Services at Canada Post Corporation. The forward to the General Glossary states that it "includes all the terms and official designations contained in our central terminology files. Its primary purpose is to standardize usage of Canada Post Terminology." This glossary has more than 6,000 entries, a figure considerably higher than the vocabulary estimate for the COPSYS manuals obtained from our corpus. Two factors account for this difference: (1) the General Glossary has a wider scope than the COPSYS manuals; (2) most of the entries in the General Glossary are multiword terms, whereas the vocabulary estimate is based on single words only. To illustrate these points, a few entries from the glossary are given below.

a. advertising ese
b. America's Cup
c. Ask About Admail
d. Canada Post Corporation belongs to all Canadians. And we're always ready to help the media tell our story.
e. "Go Canada"
f. Head Office Joint CPC-UPCE/PSAC Occupational Health and Safety Committee
g. I concur
h. Ontario Wages Act
i. Quebec City postal zone number directory
j. Sharing the Olympic Flame; Sharing the Flame
k. vibrating beeper
l. Your Postal Code Works for You.

3.3 Semantic range

The COPSYS domain is restricted to methods and procedures for processing mail. The following statement appears in the introduction to the manual:

> The COPSYS program was implemented to provide a users' guide of mail processing activities performed at the Work Centre level in mail processing facilities, (plant or office). In addition, COPSYS provides a vehicle for efficient change.

COPSYS documentation takes place at the Work Centre level and contains the following elements: general description, supervisor guidelines, safety procedures, layout, equipment, mail flow, sortation schematics, operating procedures. The domain is fairly homogeneous except for the section on safety procedures ("get a good footing; place feet shoulder width apart; keep back straight, bending at the knees; get a firm hold; keep body as upright as possible" etc.) and the sections on word processing mentioned in 3.1 above, ie, WORD PROCESSING SPECIFICATIONS and WORDSTAR FILES AND DISKETTE. The material on safety procedures is very limited, but that on word processing constitutes a significant subdomain. This does not, of course, imply that COPSYS has the field of word processing as a subdomain; only the procedures necessary for using Wordstar at a Canada Post Work Centre are included.

On the whole, the semantic range of COPSYS is very restricted.

3.4 Syntactic properties

(A) Sentence types

There are a few interrogatives in the corpus, one being a *yes/no* type and the others *wh* types. They are found mainly in the introduction, although the yes/no question("Is it originating" - with no question mark) occurs in the section PROCEDURES. Almost half of the sentences in the corpus are imperatives. These are not just parenthetical notes like "see 2.1", but longer sentences typical of instruction manuals.

(B) Topicalization

The passive is used frequently, but does not predominate. It accounts for somewhat less than half the sentences in the corpus. There are a few cases of extraposition, but no cleft

or pseudo-cleft sentences. And there are a few cases of inversion (eg, "Enclosed is a diskette with...").

(C) Clausal structure

The number of relative clauses is about average, most of them with *wh + be* deletion. Clauses introduced by subordinate conjunctions, complement clauses and abbreviated clauses are also average in number. Comparative clauses are rare. In general, sentences here contain fewer subordinate clauses than those in A.I.P.

(D) Coordination

Coordination of noun phrases is very frequent, followed by verb or verb phrase coordination and sentence coordination, in that order.

(E) Ellipsis

Article deletion occurs frequently, although not uniformly, in instructions for carrying out tasks (eg, "Place despatch bill in last bag and mark the bag label", "obtain signature of the courier/contractor on the delivery form"). There are a few subjectless sentences (eg,"Expires 30 September"). Conjunction reduction is about average.

(F) Telegraphic sentences

Telegraphic sentences occur in the corpus, but they are fewer in number and less radical in their deletions than those in weather bulletins or aviation maintenance manuals. For the most part they involve article deletion, but other types are also found (eg, "Number Bag Rack top left to bottom right", "Utilize MSC/CUS drivers available").

(G) Noun stacking

The frequency of occurrence of noun sequences in the corpus is very high. Some typical examples are:

a. Work Centre Flow Charts
b. Priority Post mail
c. mail description code
d. Change Control Procedure

e. hand cancellation
f. Street Letter Box Collections
g. main power supply machines and components
h. sealed register package envelope
i. product stream symbol
j. Letter Carrier Return Envelope
k. on the job safety instructions
I. key "cut off" time information
m. word processing (WORDSTAR) package
n. local and forward sortation and despatch schedules
o. mobile and/or stationary mail handling and storage equipment

Many of the noun sequences could probably be entered in the dictionary as idioms (eg, work centre, priority post, letter carrier, letter box, etc.); however, noun stacking is still likely to pose a problem in these texts.

(H) Parentheticals

The corpus contains a variety of parenthetical expressions. In many cases the parenthetical is, from a structural point of view, an integral part of the sentence in which it occurs, rather than something "extra" which is added to an already formed sentence. For example, in the introduction the following interrogatives have basic constituents enclosed in parentheses:

- WHO (does the job)
- WHAT (is done)
- WHEN (is it done)
- WHERE (is it done)

Certain noun phrases may be followed by abbreviations or identifiers; eg, "Acknowledgement of Receipt (AR) Card" and "Delivery Notice (Card 26)". There is some embedding of parentheticals within parentheticals: "Diskette - (containing Work Centre(s) documentation that was revised (entire Work Centre documentation))". And, of course, the usual notes to the reader, such as "(see Example 5.1.3.1)".

(I) Tense

Present tense predominates, although present perfect and simple past tense also occur.

(J) Modals and semi-auxiliaries

The corpus contains the full range of modal auxiliaries, with average frequency. Most occurrences of *will* are optative rather than future time. There are many occurrences of the semi-auxiliary *be to* and these can usually be considered as alternatives to imperative sentences; in fact, the two types often occur one after the other in lists of instructions.

(K) Sentence length

The average sentence length for the corpus is 14.2 words and 5.5% of the sentences contain 30 or more words. These figures are considerably lower than the ones for the A.I.P. corpus (23.2 words and 27.4% respectively). This is due to the large number of imperatives in COPSYS, which tend to be concise and sometimes telegraphic in style.

(L) Text structure

(i) Sentence linking. There are 18 occurrences of intersentential pronouns in the corpus. Although this figure is considerably higher than in A.I.P., it represents only about one occurrence for each 30 sentences, on the average.

(ii) Information structure. Among the imperative sentences, which constitute almost half the total number in the corpus, there is a regular correlation between information and syntatic structure: most of these sentences have verbs describing typical Work Centre activities and direct objects designating typical recipients of those actions at Canada Post Work Centres. For example:

Verb	Direct Object	Verb	Direct Object
cancel	postage stamp	sign	form, bill, sheet
date stamp	article, tag, form	prepare	form, bill, request
back stamp	article, mailings	place	item, form, impression
sort	articles, registers	indicate	office, origin, name
check off	item	initial	entry
record	item	obtain	signature
count	items	retain	copy
number	item	destroy	carbon copy
enter	number, name	maintain	record
complete	form, tag, entry	tie	registers
open	envelope	tie out	mail
slit	envelope	tray	mail
lay (flat)	envelope	write	name
face slip	bundle	affix	form
enclose	mail, bill, despatch	hand	form, item, register, lock bag

The pattern can be represented as follows:

(WORKER)	WORK CENTRE ACTIVITY	RECIPIENT OF ACTION
(Subject)	Verb	Direct Object

The same underlying relation also shows up in many non-imperative sentences throughout the corpus, in a less direct manner.

(iii) Text formatting. There are many tables in the COPSYS manual that have not been included in the corpus since they consist entirely of numerical figures and are simple in their arrangement (eg, CASE PLANS). Also omitted is a form that appears at the bottom of each page:

DOCUMENTED REVISED PLANT APPR'L COPSYS APPR'L

BY ———————————————

SIG. ———————————————

DATE ———————————————

This repeated form and the tables take up a lot of space, but have little effect on the linguistic complexity of COPSYS.

3.5 Conclusion

The COPSYS domain is very limited: mail processing activities at Canada Post Work Centres. The vocabulary is estimated at 2,000-4,000 words (lexical lemmas). The syntax is moderately complex: average in the use of subordinate clauses, high in noun stacking, average in the use of coordinate conjunctions, some interrogative sentences present in limited context, many imperatives (with frequent deletion of the definite article) and a few telegraphic sentences with deletions other than the definite article. Average sentence length for the corpus is low, at 14.2 words. Most of the imperative sentences, and many others as well, follow the pattern described in (L) (ii) above. The estimated annual volume to be translated is about 2,500,000 words. Translation turnaround time is currently one week, although a desire has been expressed to shorten that time. Tolerance of rigid style is very good (establishment of norms for writing has been considered as one way of coping with the variations arising from the large number of authors involved in updating COPSYS manuals throughout Canada Post). Compared with A.I.P., COPSYS has a more restricted semantic domain, a smaller vocabulary and less syntactic complexity.

However, COPSYS cannot be rated as low in overall linguistic complexity; it is moderately complex. In particular, the high incidence of noun stacking could prove to be a very serious problem for automatic parsing in this domain, although, as pointed out in 3.4(G), the extent of the problem may be reduced somewhat by entering certain noun strings in the dictionary as single lexical items. On balance, machine translation appears to be feasible for the COPSYS manuals.

4. CENSUS AND INTERCENSAL STUDIES

This domain will be referred to as CENSUS. It contains not only census data, but interpretation and analysis of the data as well, and it is one of the eight categories into which Statistics Canada publications are divided (General; Primary Industries; Manufacturing; Transportation, Communications and Utilities; Commerce, Construction, Finance and Prices; Employment, Unemployment and Labour Income; Education, Culture, Health and Welfare; Census and Intercensal Studies).

A census is taken every five years and the annual volume of census related text to be translated varies considerably during that interval. The volume is greatest in the year preceding the census because of preparatory work, questionnaires, etc. For 1988-89 the volume was 500,000 - 700,000 words. For 1990-91 the projected volume is 3.6 million words. The total volume for all STATSCAN publications in 1988-89 was 6.2 million words. The translation turnaround time for census related texts varies from about one to six weeks.

4.1 The corpus

The corpus consists of 12,009 words of text (as counted by FATRAS) taken from the following publications:

- QUARTERLY DEMOGRAPHIC STATISTICS
- PROVINCIAL SERIES: QUEBEC
- POPULATION AND DWELLING COUNTS
- SELECTED CHARACTERISTICS: TORONTO
- CENSUS OF AGRICULTURE: LIVESTOCK AND POULTRY
- SELECTED CHARACTERISTICS: BRITISH COLUMBIA
- PRODUCTS AND SERVICES
- CANADA'S NATIVE PEOPLE
- CURRENT DEMOGRAPHIC ANALYSIS:
  - population structure
  - marriage and divorce
  - mortality
  - migration

The publications in Census and Intercensal Studies include a very large number of tables and charts. However, since numerical entries and lists of place names do not add significantly to the linguistic complexity of the text, tables and charts are not included in the corpus.

4.2 Size of vocabulary

The corpus was divided into 12 segments of about 1,000 words each for the vocabulary study. Table 3 shows the number of words in each segment and the lexical growth rate there (ie, the number of lemmas introduced per thousand words of text). Figure 5, page 30, is the lexical growth rate curve obtained from this data.

SEGMENT	WORDS	LGR	SEGMENT	WORDS	LGR
1	1,000	216	7	1,000	114
2	1,000	183	8	1,001	79
3	1,002	181	9	1,001	73
4	1,003	159	10	1,000	76
5	1,002	101	11	999	80
6	1,001	93	12	1,000	73

TABLE 3

There is a steady decline in the lexical growth rate from a value of 216 in the first segment to 93 in the sixth. The trend is then reversed as the value increases to 114 in the seventh segment. The value falls off again, with only a minor reversal in the tenth and eleventh segments.

In Figure 6, page 31, the curve for CENSUS is compared with those for A.I.P. and COPSYS. After 7,000 words of text the number of lexical lemmas in CENSUS is greater than in A.I.P. or COPSYS. After about 9,000 words the growth rates of CENSUS and A.I.P. (ie, the slopes of their curves) are roughly the same. The vocabulary required for the CENSUS domain can be expected to exceed 5,000 lexical lemmas, exclusive of place names.

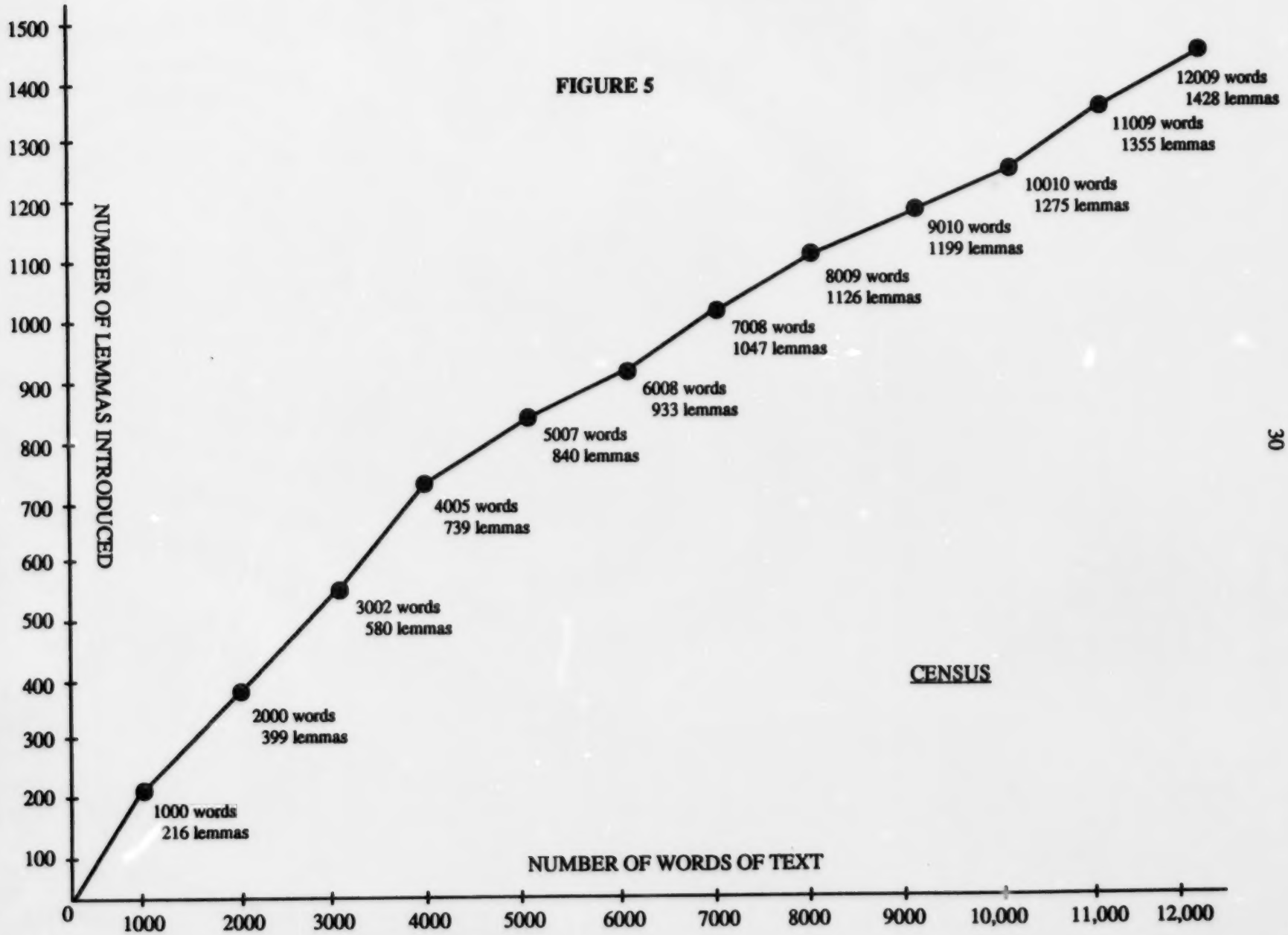
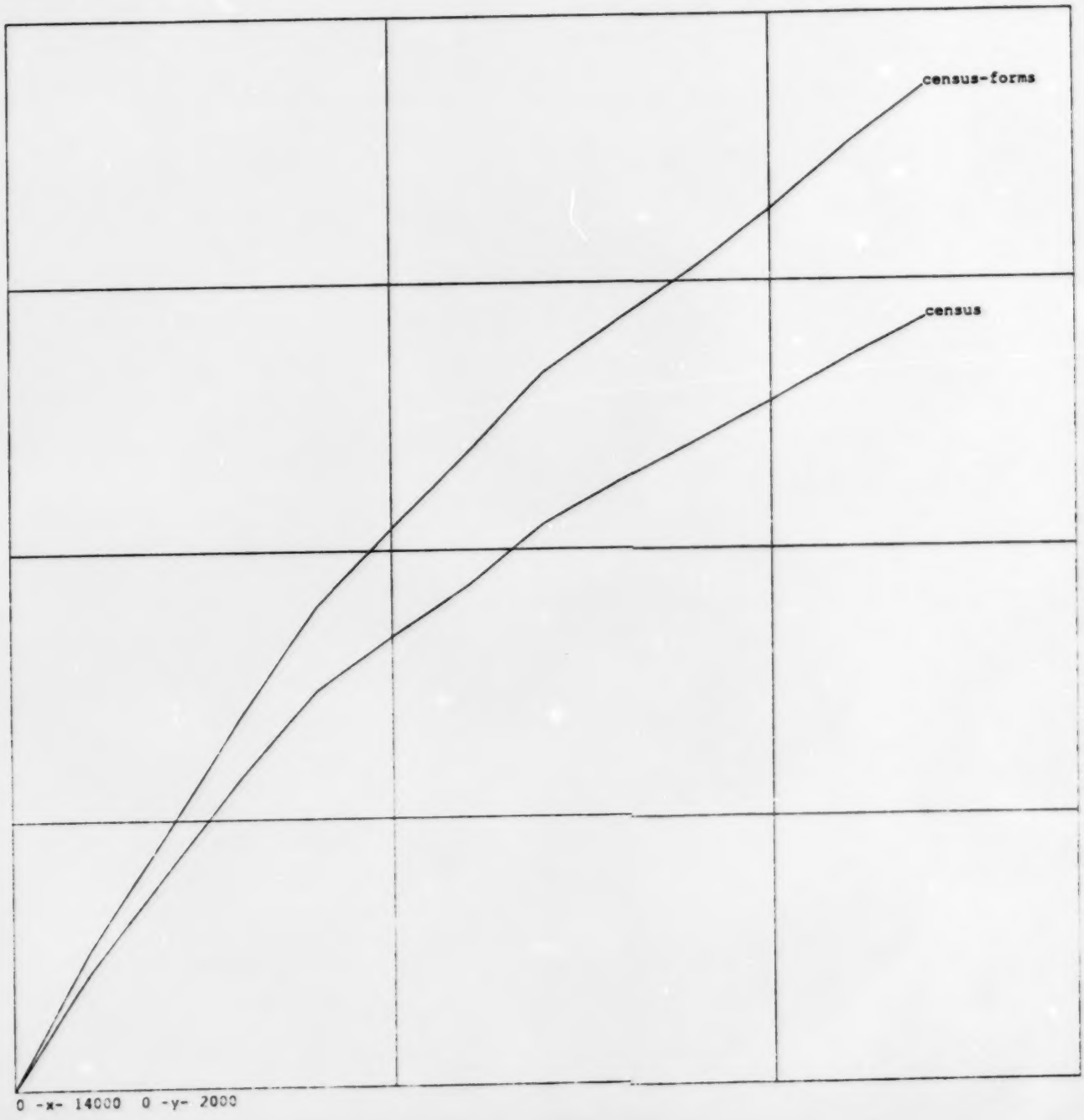


FIGURE 5
NUMBER OF LEMMAS INTRODUCED
NUMBER OF WORDS OF TEXT
CENSUS
1000 words
216 lemmas
2000 words
399 lemmas
3002 words
580 lemmas
4005 words
739 lemmas
5007 words
840 lemmas
6008 words
933 lemmas
7008 words
1047 lemmas
8009 words
1126 lemmas
9010 words
1199 lemmas
10010 words
1275 lemmas
11009 words
1355 lemmas
12009 words
1428 lemmas
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10,000
11,000
12,000

Figure 6

Figure 7



As explained in section 1.3, lexical *lemmas* rather than lexical *forms* have formed the basis for these vocabulary studies. The present corpus, at more than 12,000 words, is sufficiently large to give a good idea of the difference in results obtained by the two approaches. Figure 7, page 32, shows the relation between the lexical growth rate curves in terms of lemmas and forms, using the CENSUS corpus.

4.3 Semantic range

CENSUS covers a wide range of topics in addition to population counts with geographic distribution and changes over time. It also includes information on:

- birth rates
- death rates
- life expectancy
- marriage rates
- divorce rates
- migration (international and interprovincial)
- linguistic mobility
- native people (Indians, Inuit)
- occupations and labour market areas
- dwelling places (private, collective, structural type)
- coverage errors (missing a person or household, or counting them more than once)
- breakdown of population by:

mother tongue	religion
official language	place of birth
home language	schooling
ethnic origin	age, sex

- census of agriculture (livestock and poultry, etc.)

It is important to remember that the CENSUS texts include not only statistical reports on these subjects, but analysis of the results as well. The semantic range of CENSUS is then rather broad, and this is reflected in the size of vocabulary required for the domain.

4.4 Syntactic properties

(A) Sentence types

There are three interrogatives in the corpus, all of the *wh* type. They occur as section headings in the publication **Canada's Native People** (eg, "How large are the main native groups and where do they live?"). The only imperative is a parenthetical note ("see Appendix A-5"); this is to be expected, since the texts in this corpus are not sets of instructions, but reports in which results are discussed and analyzed.

(B) Topicalization

Nearly forty percent (40%) of the sentences in the corpus are in the passive. Extraposition, cleft and pseudo-cleft sentences are very rare. There are several cases of inversion, eg, "Also included are private dwellings".

(C) Clausal structure

Relative clauses occur with high frequency, especially those with *wh* + *be* deletion. Complement clauses and clauses introduced by subordinate conjunctions are about average in number. There are several occurrences of abbreviated clauses (eg, 'when applicable", wherever feasible"), but they are much less frequent here than in COPSYS or A.I.P.

(D) Coordination

The frequency of coordination is very high, with noun phrase coordination accounting for almost sixty percent of the total number of occurrences, followed by verb (or verb phrase) and sentence coordination, in that order. Conjuncts that do not belong to any recognized category are also found (eg, "impossible to determine specific *age at* and *year of* immigration").

(E) Ellipsis

There are more than thirty subjectless sentences in the corpus. They occur in sections where definitions are stated, the omitted subject being the term which is defined (eg, 'Age. Refers to age at last birthday"). This is the usual format, although the subject is not always omitted ("Place of birth. For persons born in Canada, place of birth refers to the specific

province or territory of birth"). Conjunction reduction is a major source of ellipsis, due to the large number of coordinate conjunctions in the corpus. There are just a few occurrences of article deletion, confined to specific contexts such as definitions.

(F) Telegraphic sentences

This type of text makes very little use of telegraphic sentences. The following occur in explaining the use of the symbols ".." and "- - ":

.. figures not available

- - amount too small to be expressed

(G) Noun stacking

Noun stacking is less frequent than in either A.I.P. or COPSYS and there are usually only two nouns involved. Quite a few of these nominal compounds are designated by their initials, as indicated in parentheses in the following examples:

birth rate
growth rate
age group
life expectancy
census tract (CT)
census division (CD)
census agglomeration (CA)
enumeration area (EA)
census metropolitan area (CMA)
population count
population density
household income
sample data base
Indian reserve
Status Indians
home language
mother tongue

It is likely that many such nominal compounds could be entered in the dictionary as idioms. Noun stacking poses less of a problem in the CENSUS domain than in A.I.P. or COPSYS, although this may not be the case for STATSCAN publications in general.

(H) Parentheticals

The frequency of occurrence of parentheticals is average, but there is considerable variety in the type of parenthetical expression and its position in the sentence. A typical use is the placement of numerical expressions following quantitative statements:

> This amounts to a 15% increase in annual growth compared to last year (227,500).
>
> The average period for each cycle has been eight or nine years, with peaks occurring in 1951 (190,000), 1957 (200,000), 1967 and 1974 (220,000), and in 1980 (a "low" peak of only 140,000) (Chart 12).

In addition to parentheses dashes are also used to set off parenthetical expressions:

> Among males, only two other provinces - Newfoundland and Alberta - have a rate of first marriage lower than the Canadian average.
>
> ... the most widely known indicators - expectation of life at different ages, and especially at birth - are taken.

(I) Tense

Present tense occurs most frequently, although past tense is also used a great deal.

(J) Modals and semi-auxiliaries

The corpus contains the full range of modal auxiliaries, but they are used less here than in A.1.P. or COPSYS. The use of semi-auxiliaries is rare (eg, "only the respondent's paternal ancestry was to be reported").

(K) Sentence length

The average sentence length for the corpus is 19.6 words and 11.7% of the sentences contain 30 or more words. The biggest contributor to this high average is the publication *Current Demographic Analysis,* which has many very long sentences. These figures are higher than the ones for COPSYS (14.2 words and 5.5% respectively), but lower than those for A.I.P. (23.2 words and 27.4% respectively).

(L) Text structure

(i) Sentence linking. There are 28 occurrences of intersentential pronouns in the corpus. This represents about one occurrence for each 20 sentences, a figure slightly higher than that for COPSYS.

(ii) Information structure. The sentences, or clauses, of the text are of two types: (1) core sentences, which give quantitative information (population counts, birth rates, immigration figures, etc., and changes in these values since earlier censuses); (2) non-core sentences, which interpret and analyze the quantitative information, or which give background information on the categories relevant to the CENSUS domain. The core sentences in CENSUS are of the same general form as those in other STATSCAN publications (see *Machine Translatability of STATSCAN Publications*, J. Lehrberger, CWARC 1988, section III: Summary and Conclusion). Non-core sentences predominate in introductory sections, definitions, discussions of background material, methodology, etc.

(iii) Text formatting. The CENSUS publications contain a very large number of tables with many place names and numerical entries. These have not been included in the corpus since they have little effect on the linguistic complexity of CENSUS. Also, there are some charts containing small amounts of text arranged in such a manner that the translation unit may not be clear to the computer; however, such cases are few in number and involve very little text. Otherwise, text formatting does not appear to present any serious problems.

4.5 Conclusion

The CENSUS domain covers a fairly wide range of topics, as pointed out in 4.3. Its semantic range is greater than that of COPSYS. The vocabulary required for the entire CENSUS domain can be expected to exceed 5,000 words (lexical lemmas), exclusive of place names. The syntax is complex, with many subordinate clauses and much use of coordinate conjunctions. Interrogatives, imperatives and telegraphic sentences are rare in the corpus, confined to specific contexts. Noun stacking is less of a problem than in either A.I.P. or COPSYS, usually involving only two nouns. Core sentences form a significant component of the text. They are not necessarily simple sentences, but follow certain patterns which are well correlated with information content. Average sentence length is a little high at just under 20 words, although not as high as in A.I.P. The annual volume of text to be translated varies from about half a million to 3.6 million words during the interval from one census to the next.

CENSUS is a complex domain. It ranks lower than COPSYS as a candidate for machine translation, although the prevalence of core sentences suggests that it should not be rejected as a candidate altogether. Furthermore, the fact that these core sentences are of the same general type as those found throughout most STATSCAN publications, suggests that a system designed to translate census related material might later be extended to cover the broader field.

Pour plus de détails, veuillez communiquer avec :

Le Centre canadien de recherche sur l'informatisation du travail
1575, boulevard Chomedey
Laval (Québec)
H7V 2X2
(514) 682-3400

For more information, please contact:

Canadian Workplace Automation Research Centre
1575 Chomedey Blvd.
Laval, Quebec
H7V 2X2
(514) 682-3400